GitHub Copilot AI pair programmer: Asset or Liability?

Abstract

Automatic program synthesis is a long-lasting dream in software engineering. Recently, a promising Deep Learning (DL) based solution, called Copilot, has been proposed by OpenAI and Microsoft as an industrial product. Although some studies evaluate the correctness of Copilot solutions and report its issues, more empirical evaluations are necessary to understand how developers can benefit from it effectively. In this paper, we study the capabilities of Copilot in two different programming tasks: (i) generating (and reproducing) correct and efficient solutions for fundamental algorithmic problems, and (ii) comparing Copilot's proposed solutions with those of human programmers on a set of programming tasks. For the former, we assess the performance and functionality of Copilot in solving selected fundamental problems in computer science, like sorting and implementing data structures. In the latter, a dataset of programming problems with human-provided solutions is used. The results show that Copilot is capable of providing solutions for almost all fundamental algorithmic problems, however, some solutions are buggy and non-reproducible. Moreover, Copilot has some difficulties in combining multiple methods to generate a solution. Comparing Copilot to humans, our results show that the correct ratio of humans' solutions is greater than Copilot's suggestions, while the buggy solutions generated by Copilot require less effort to be repaired.